BIMM-143: INTRODUCTION TO BIOINFORMATICS

The find-a-gene project assignment
https://bioboot.github.io/bimm143_S20/
Dr. Barry Grant

Overview:

The find-a-gene project is a required assignment for BIMM-143. You should prepare a written report in **PDF** format that has responses to each question labeled **[Q1] - [Q10]** below. You may wish to consult the scoring rubric at the end of this document and the example report provided online.

The objective with this assignment is for you to demonstrate your grasp of database searching, sequence analysis, structure analysis and the R environment that we have covered in class.

Due Date:

Your responses to questions Q1-Q4 are due at the beginning of class **Tuesday May 5th** (05/05/20) at 12pm San Diego time. Note that these answers can be obtained very quickly (at best within 10 or 15 minutes), so if you don't succeed at first, just keep trying.

The complete assignment, including responses to all questions, is due **Friday June 5th** (06/05/20) at 12pm San Diego time.

Submission instructions:

Your report formatted as a **PDF document** should be uploaded to ***GradeScope***. Please make sure to include your UCSD email and PID number on the first page.

Be sure to include your UCSD email and PID number on the first page of your report.

Submit your preliminary report with answers to Q1-Q4 as soon as you can so we can determine if you have found a novel gene. Submit this preliminary report as one document with screen shots of the results inserted appropriately.

See the demonstration report linked to on the course website for an example of format. I will email you my decision; proceed with subsequent questions only after we are sure you have found a novel gene.

For the final report add your results for Q5-Q10 to the preliminary report and submit the final document containing your results for all questions - **Please do not send only Q5-Q10 answers as the final report**.

Questions:

[Q1] Tell me the name of a protein you are interested in. Include the species and the accession number. This can be a human protein or a protein from any other species as long as it's function is known.

If you do not have a favorite protein, select human RBP4 or KIF11. Do not use beta globin as this is in the worked example report that I provide you with online.

Name: human RBP4 (retinol-binding protein 4), it binds and transports vitamin A, retinol, from the liver to peripheral tissues
Species: homo sapien
Protein accession # : NP_006735.2

[Q2] Perform a BLAST search against a DNA database, such as a database consisting of genomic DNA or ESTs. The BLAST server can be at NCBI or elsewhere. Include details of the BLAST method used, database searched and any limits applied (e.g. Organism).

NCBI tblastn
Database : expressed sequences tags (est), No organism specified

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
ILLUMIGEN_MCQ_57772 Katze_MNLV Macaca nemestrina cDNA clone IBIUW:33878 5' similar to Bases 5 to 67...	Macaca nemestrina	399	399	100%	2e-140	94.03%	770	DR774086.1
13080 Full Length cDNA from the Mammalian Gene Collection Homo sapiens cDNA 5' similar to BC020633, mRN...	Homo sapiens	399	399	99%	3e-140	95.00%	750	EL735622.1
ILLUMIGEN_MCQ_56576 Katze_MMLV Macaca mulatta cDNA clone IBIUW:30796 5' similar to Bases 4 to 805 hi...	Macaca mulatta	399	399	100%	7e-140	94.53%	830	DR773728.1
ILLUMIGEN_MCQ_53621 Katze_MNLV Macaca nemestrina cDNA clone IBIUW:32842 5' similar to Bases 5 to 86...	Macaca nemestrina	400	400	100%	9e-140	94.53%	961	DR772055.1
DC644846 macaque kidney cDNA library QreB Macaca fascicularis cDNA clone QreB-20384 5', mRNA sequence	Macaca fascicularis	399	399	100%	1e-139	94.53%	875	DC644846.1
ILLUMIGEN_MCQ_58883 Katze_MMLV Macaca mulatta cDNA clone IBIUW:33106 5' similar to Bases 5 to 789 hi...	Macaca mulatta	399	399	100%	1e-139	94.53%	844	DR774611.1
DC626304 macaque liver cDNA library QlvC Macaca fascicularis cDNA clone QlvC-20066 5', mRNA sequence	Macaca fascicularis	399	399	100%	1e-139	94.53%	886	DC626304.1
DC629221 macaque liver cDNA library QlvC Macaca fascicularis cDNA clone QlvC-30039 5', mRNA sequence	Macaca fascicularis	399	399	100%	2e-139	94.53%	936	DC629221.1
ILLUMIGEN_MCQ_49216 Katze_MNLV Macaca nemestrina cDNA clone IBIUW:19625 5' similar to Bases 484 to ...	Macaca nemestrina	399	399	100%	2e-139	94.53%	956	CO580090.1
DC622649 macaque liver cDNA library QlvC Macaca fascicularis cDNA clone QlvC-07358 5', mRNA sequence	Macaca fascicularis	399	399	100%	2e-139	94.53%	943	DC622649.1
DC626145 macaque liver cDNA library QlvC Macaca fascicularis cDNA clone QlvC-19228 5', mRNA sequence	Macaca fascicularis	399	399	100%	3e-139	94.53%	956	DC626145.1
DC621124 macaque liver cDNA library QlvC Macaca fascicularis cDNA clone QlvC-02335 5', mRNA sequence	Macaca fascicularis	399	399	100%	3e-139	94.53%	914	DC621124.1
DC622141 macaque liver cDNA library QlvC Macaca fascicularis cDNA clone QlvC-06076 5', mRNA sequence	Macaca fascicularis	399	399	100%	4e-139	94.53%	987	DC622141.1
ILLUMIGEN_MCQ_56519 Katze_MMLV Macaca mulatta cDNA clone IBIUW:33106 5' similar to Bases 5 to 848 hi...	Macaca mulatta	397	397	100%	4e-139	94.53%	891	DR773726.1
DC639849 macaque kidney cDNA library QreB Macaca fascicularis cDNA clone QreB-27010 5', mRNA sequence	Macaca fascicularis	398	398	100%	5e-139	94.53%	933	DC639849.1
DC621287 macaque liver cDNA library QlvC Macaca fascicularis cDNA clone QlvC-03161 5', mRNA sequence	Macaca fascicularis	399	399	100%	5e-139	94.53%	1006	DC621287.1
DC626243 macaque liver cDNA library QlvC Macaca fascicularis cDNA clone QlvC-19364 5', mRNA sequence	Macaca fascicularis	399	399	100%	5e-139	94.53%	985	DC626243.1
DC622429 macaque liver cDNA library QlvC Macaca fascicularis cDNA clone QlvC-07059 5', mRNA sequence	Macaca fascicularis	397	397	100%	1e-138	94.03%	949	DC622429.1

Download GenBank Graphics Next Previous Descriptions

ILLUMIGEN_MCQ_57772 Katze_MNLV Macaca nemestrina cDNA clone IBIUW:33878 5' similar to Bases 5 to 670 highly similar to human RBP4 (Hs.50223), mRNA sequence

Sequence ID: DR774086.1 Length: 770 Number of Matches: 1

Range 1: 82 to 684 GenBank Graphics Next Match Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
399 bits(1026)	2e-140	Compositional matrix adjust.	197/201(98%)	201/201(100%)	0/201(0%)	+1

```
Query  1    MKWVWallllaalGSGRAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIV  60
            MKWVWALLLLAALGSGRAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIV
Sbjct  82   MKWVWALLLLAALGSGRAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIV  261

Query  61   AEFSVDETGQMSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGND  120
            AEFSVDETGQMSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGND
Sbjct  262  AEFSVDETGQMSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGND  441

Query  121  DHWIVDTDYDTYAVQYSCRLLNLDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQEELCLA  180
            DHWI+DTDYDTYAVQYSCRLLNLDGTCADSYSFVFSRDPNGLPPEAQ+IVRQRQEELCLA
Sbjct  442  DHWIIDTDYDTYAVQYSCRLLNLDGTCADSYSFVFSRDPNGLPPEAQRIVRQRQEELCLA  621

Query  181  RQYRLIVHNGYCDGRSERNLL  201
            RQYRLIVHNGYCDG+SE+NLL
Sbjct  622  RQYRLIVHNGYCDGKSEKNLL  684
```

Chosen Match
Accession # : DR774086.1
A 770 cDNA clone from Macaca nemestrina

[Q3] Gather information about this "novel" **protein**. At a minimum, show me the protein sequence of the "novel" protein as displayed in your BLAST results from [Q2] as FASTA format (you can copy and paste the aligned sequence subject lines from your BLAST result page if necessary) or translate your novel DNA sequence using a tool called EMBOSS Transeq at the EBI. Don't forget to translate all six reading frames; the ORF (open reading frame) is likely to be the longest sequence without a stop codon. It may not start with a methionine if you don't have the complete coding region. Make sure the sequence you provide includes a header/subject line and is in traditional FASTA format.

>M. nemestrina protein
MKWVWALLLLAALGSGRAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQMSATAK
GRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGNDDHWIIDTDYDTYAVQYSCRLLNLDGTCADSY
SFVFSRDPNGLPPEAQRIVRQRQEELCLARQYRLIVHNGYCDGKSEKNLL

Here, tell me the name of the novel protein, and the species from which it derives. It is very unlikely (but still definitely possible) that you will find a novel gene from an organism such as *S. cerevisiae*, human or mouse, because those genomes have already been thoroughly annotated. It is more likely that you will discover a new gene in a genome that is currently being sequenced, such as bacteria or plants or protozoa.

Name : Macaca RBP4
Species : Macaca nemestrina
Animalia; Chordata; Mammalia; Primates; Haplorhini; Simiiformes; Cercopithecidae; Macaca; Nemestrina

[Q4] Prove that this gene, and its corresponding protein, are novel. For the purposes of this project, "novel" is defined as follows. Take the protein sequence (your answer to [Q3]), and use it as a query in a blastp search of the nr database at NCBI.

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
PREDICTED: retinol-binding protein 4 [Macaca fascicularis]	Macaca fasci…	418	418	100%	1e-147	99.00%	201	XP_005566031.1
retinol-binding protein 4 isoform X1 [Hylobates moloch]	Hylobates m…	418	418	100%	2e-147	98.51%	214	XP_031992464.1
retinol-binding protein 4 [Papio anubis]	Papio anubis	417	417	100%	3e-147	98.51%	201	XP_003904062.1
retinol-binding protein 4 precursor [Pan troglodytes]	Pan troglodytes	416	416	100%	5e-147	98.01%	201	NP_001038960.1
retinol-binding protein 4 [Nomascus leucogenys]	Nomascus le…	416	416	100%	8e-147	98.01%	201	XP_003255281.1
PREDICTED: retinol-binding protein 4 [Rhinopithecus bieti]	Rhinopithecu…	416	416	100%	9e-147	98.01%	201	XP_017732256.1
Retinol binding protein 4, plasma [Homo sapiens]	Homo sapiens	415	415	100%	2e-146	97.51%	201	AAH20633.1

A blastp search against NR database yielded a top hit from Macaca fascicularis
The top result is to a protein from Macaca fascicularis

```
Score          Expect  Method                          Identities      Positives       Gaps
418 bits(1075) 1e-147  Compositional matrix adjust.    199/201(99%)    201/201(100%)   0/201(0%)

Query  1    MKWVWALLLLAALGSGRAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIV  60
            MKWVWALLLLAALGSGRAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIV
Sbjct  1    MKWVWALLLLAALGSGRAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIV  60

Query  61   AEFSVDETGQMSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGND  120
            AEFSVDETGQMSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGND
Sbjct  61   AEFSVDETGQMSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGND  120

Query  121  DHWIIDTDYDTYAVQYSCRLLNLDGTCADSYSFVFSRDPNGLPPEAQRIVRQRQEELCLA  180
            DHWIIDTDYDTYAVQYSCRLLNLDGTCADSYSFVFSRDPNGLPPEAQRIVRQRQEELCLA
Sbjct  121  DHWIIDTDYDTYAVQYSCRLLNLDGTCADSYSFVFSRDPNGLPPEAQRIVRQRQEELCLA  180

Query  181  RQYRLIVHNGYCDGKSEKNLL  201
            RQYRLIVHNGYCDG+SE+NLL
Sbjct  181  RQYRLIVHNGYCDGRSERNLL  201
```

[Q5] Generate a multiple sequence alignment with your novel protein, your original query protein, and a group of other members of this family from different species. A typical number of proteins to use in a multiple sequence alignment for this assignment purpose is a minimum of 5 and a maximum of 20 - although the exact number is up to you. Include the multiple sequence alignment in your report. Use Courier font with a size appropriate to fit page width.

Side-note: Indicate your sequence in the alignment by choosing an appropriate name for each sequence in the input unaligned sequence file (i.e. edit the sequence file so that the species, or short common, names (rather than accession numbers) display in the output alignment and in the subsequent answers below). The goal in this step is to create an interesting an alignment for building a phylogenetic tree that illustrates species divergence.

>NP_006735.2 retinol-binding protein 4 isoform a precursor [Homo sapiens]
MKWVWALLLLAALGSGRAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQ
MSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYAVQYSCRL
LNLDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQEELCLARQYRLIVHNGYCDGRSERNLL
>XP_005566031.1 PREDICTED: retinol-binding protein 4 [Macaca fascicularis]
MKWVWALLLLAALGSGRAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQ
MSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGNDDHWIIDTDYDTYAVQYSCRL
LNLDGTCADSYSFVFSRDPNGLPPEAQRIVRQRQEELCLARQYRLIVHNGYCDGRSERNLL
>XP_031992464.1 retinol-binding protein 4 isoform X1 [Hylobates moloch]
MEASLPQGGFLGKMKWVWALLLLAALGSGRAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQD
NIVAEFSVDETGQMSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGNDDHWIIDT
DYDTYAVQYSCRLLNLDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQEELCLARQYRLIVHNGYCDGRSE
RNLL

>NP_001038960.1 retinol-binding protein 4 precursor [Pan troglodytes]
MKWVWALLLLAALGSGRAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQ
MSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGNDDHWIVDTDYDTYAVQYSCRL
LNLDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQEELCLARQYRLIVHNGYCDGRSERNLL
>XP_003904062.1 retinol-binding protein 4 [Papio anubis]
MKWVWALLLLAALGSGRAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQ
MSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGNDDHWIIDTDYDTYAVQYSCRL
LNLDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQEELCLARQYRLIVHNGYCDGRSERNLL
>XP_017732256.1 PREDICTED: retinol-binding protein 4 [Rhinopithecus bieti]
MKWMWALLLLAALGSGRAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAEFSVDETGQ
MSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGNDDHWIIDTDYDTYAVQYSCRL
LNLDGTCADSYSFVFSRDPNGLPPEAQKIVRQRQEELCLARQYRLIVHNGYCDGRSERNLL

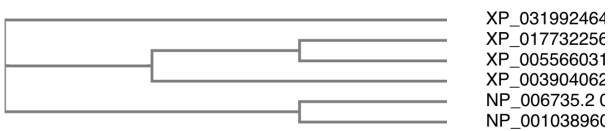
CLUSTAL multiple sequence alignment by MUSCLE (3.8)

```
M.              ----------MKWVWALLLLAALGSGRAERDCRVSSFRVKENFDKARFSGTWYAMAK
Hylobates       MEASLPQGGFLGKMKWVWALLLLAALGSGRAERDCRVSSFRVKENFDKARFSGTWYAMAK
Macaca          ----------MKWVWALLLLAALGSGRAERDCRVSSFRVKENFDKARFSGTWYAMAK
NP_006735.2     ----------MKWVWALLLLAALGSGRAERDCRVSSFRVKENFDKARFSGTWYAMAK
Papio           ----------MKWVWALLLLAALGSGRAERDCRVSSFRVKENFDKARFSGTWYAMAK
                          **********************************************

M.              KDPEGLFLQDNIVAEFSVDETGQMSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMK
Hylobates       KDPEGLFLQDNIVAEFSVDETGQMSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMK
Macaca          KDPEGLFLQDNIVAEFSVDETGQMSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMK
NP_006735.2     KDPEGLFLQDNIVAEFSVDETGQMSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMK
Papio           KDPEGLFLQDNIVAEFSVDETGQMSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMK
                ************************************************************

M.              YWGVASFLQKGNDDHWIIDTDYDTYAVQYSCRLLNLDGTCADSYSFVFSRDPNGLPPEAQ
Hylobates       YWGVASFLQKGNDDHWIIDTDYDTYAVQYSCRLLNLDGTCADSYSFVFSRDPNGLPPEAQ
Macaca          YWGVASFLQKGNDDHWIIDTDYDTYAVQYSCRLLNLDGTCADSYSFVFSRDPNGLPPEAQ
NP_006735.2     YWGVASFLQKGNDDHWIVDTDYDTYAVQYSCRLLNLDGTCADSYSFVFSRDPNGLPPEAQ
Papio           YWGVASFLQKGNDDHWIIDTDYDTYAVQYSCRLLNLDGTCADSYSFVFSRDPNGLPPEAQ
                ****************.*******************************************

M.              RIVRQRQEELCLARQYRLIVHNGYCDGKSEKNLL
Hylobates       KIVRQRQEELCLARQYRLIVHNGYCDGRSERNLL
Macaca          RIVRQRQEELCLARQYRLIVHNGYCDGRSERNLL
NP_006735.2     KIVRQRQEELCLARQYRLIVHNGYCDGRSERNLL
Papio           KIVRQRQEELCLARQYRLIVHNGYCDGRSERNLL
                .*************************.**.***
```

Question 6 Cladogram

Phylogram

Branch length: Cladogram Real

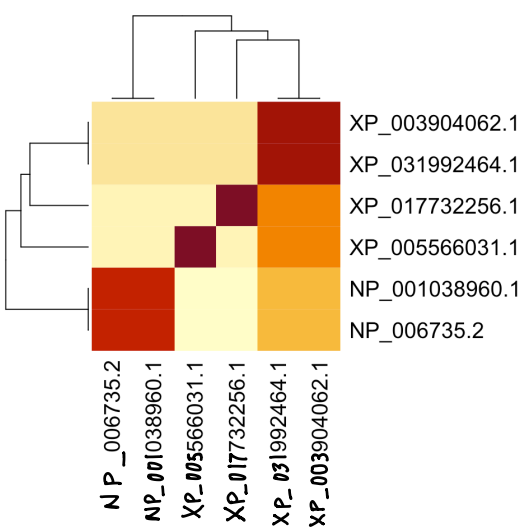


Phylogenetic Tree

View Phylogenetic Tree File

```
(
XP_031992464.1:0.00000,
(
(
XP_017732256.1:0.00498,
XP_005566031.1:0.00498)
:0.00000,
XP_003904062.1:0.00000)
:0.00000,
(
NP_006735.2:0.00000,
NP_001038960.1:0.00000)
:0.00497);
```

Question 7 Heatmap



[Q6] Create a phylogenetic tree, using either a parsimony or distance-based approach. Bootstrapping and tree rooting are optional. Use “simple phylogeny” online from the EBI or any respected phylogeny program (such as MEGA, PAUP, or Phylip). Paste an image of your Cladogram or tree output in your report.

[Q7] Generate a sequence identity based **heatmap** of your aligned sequences using R.

If necessary convert your sequence alignment to the ubiquitous FASTA format (Seaview can read in clustal format and “Save as” FASTA format for example). Read this FASTA format alignment into R with the help of functions in the **Bio3D package**. Calculate a sequence identity matrix (again using a function within the Bio3D package). Then generate a heatmap plot and add to your report. Do make sure your labels are visible and not cut at the figure margins.

[Q8] Using R/Bio3D (or an online blast server if you prefer), search the main protein structure database for the most similar atomic resolution structures to your aligned sequences.

List the top 3 *unique* hits (i.e. not hits representing different chains from the same structure) along with their Evalue and sequence identity to your query. Please also add annotation details of these structures. For example include the annotation terms PDB identifier (structureId), Method used to solve the structure (experimentalTechnique), resolution (resolution), and source organism (source).

HINT: You can use a single sequence from your alignment or generate a consensus sequence from your alignment using the Bio3D function consensus(). The Bio3D functions blast.pdb(), plot.blast() and pdb.annotate() are likely to be of most relevance for completing this task. Note that the results of blast.pdb() contain the hits PDB identifier (or pdb.id) as well as Evalue and identity. The results of pdb.annotate() contain the other annotation terms noted above.

Note that if your consensus sequence has lots of gap positions then it will be better to use an original sequence from the alignment for your search of the PDB. In this case you could chose the sequence with the highest identity to all others in your alignment by calculating the row-wise maximum from your sequence identity matrix.

[Q9] Generate a molecular figure of one of your identified PDB structures using the **NGL viewer** online (or **VMD/PyMol)**. You can optionally highlight conserved residues that are likely to be functional. Please use a white or transparent background for your figure (i.e. not the default black).

Based on sequence similarity. How likely is this structure to be similar to your “novel” protein?

[Q10] Perform a “Target” search of ChEMBEL (https://www.ebi.ac.uk/chembl/) with your novel sequence. Are there any **Target Associated Assays** and **ligand efficiency data** reported that may be useful starting points for exploring potential inhibition of your novel protein?

Scoring Rubric:

[45 total points available]

Q1 (4 points)

Protein name	1
Species	1
Accession number	1
Function known	1

Q2 (6 points)

Blast method	1
Database searched	1
Limits applied	1
Search output list (top hits)	1
Alignment of choice	1
Evalue and other alignment stats	1

Q3 (3 points)

Protein sequence of choice matches Subject above	1
Name in header	1
Species	1

Q4 (3 point)

Blastp output list with identities & Evalue	1
Top alignment shown with alignment statistics	1

Results indicates a “novel” gene found	1

Q5 (3 points)

MSA labeled with useful names	1
MSA trimmed appropriately (i.e. no gap overhangs)	1
Pasted MSA fits report page width (i.e. font, format)	1

Q6 (1 point)

Figure illustrates sequence clustering pattern	1

Q7 (10 points)

Heatmap figure included in report	5
Heatmap is legible (i.e. no labels obscured)	5

Q8 (10 points)

PDB identifiers from multiple species reported	5
Annotation of PDB source, resolution and technique	4
Annotation of Evalue and Sequence Identity	1

Q9 (4 points)

Structure figure provided	2
Uses white background for molecular figure	1
Figure of high resolution (i.e. not just snapshot)	1

Q10 (1 point)

Evidence of ChEMBEL searches	1

Find A Gene

Hayoung A15531571

12/2/2021

Load in Bio3D

```
library(bio3d)
library(plyr)
library(ggplot2)
```

Q7. Generate a sequence identity based heatmap

Let's read our fasta file into R!

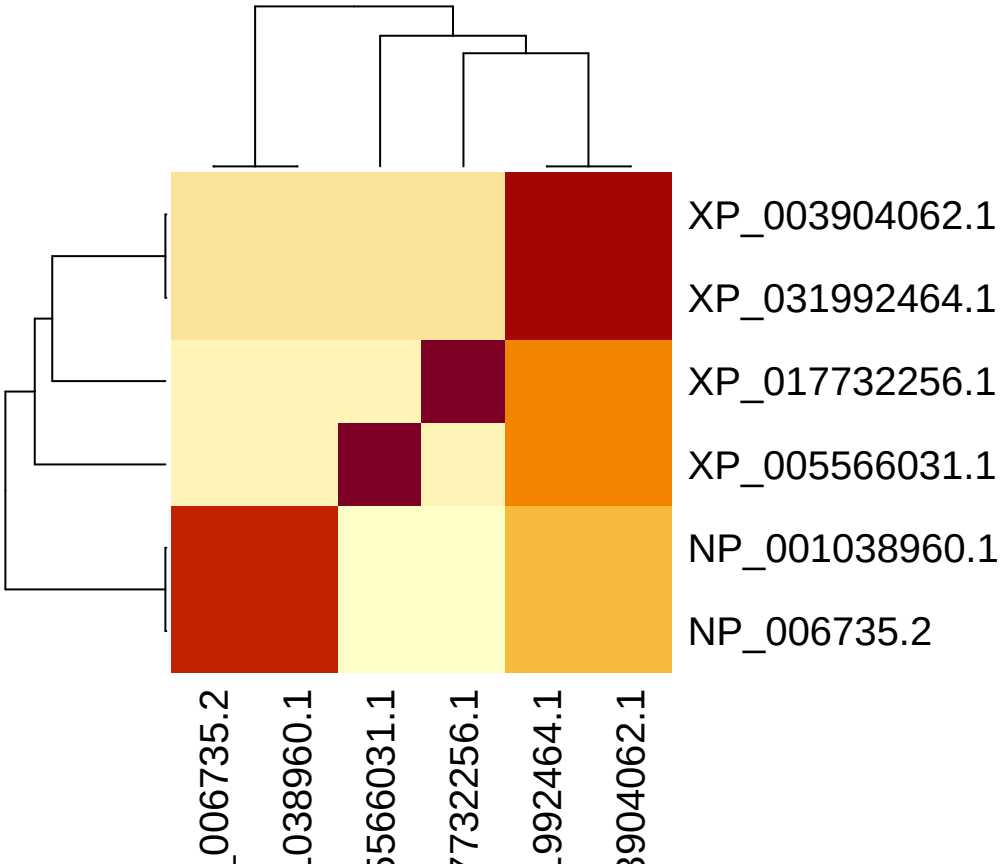
```
fast <- read.fasta("muscle-I20211203-024209-0985-8072325-p2m.clw.fst")
clust <- read.csv("muscle-I20211203-024209-0985-8072325-p2m.clw")
```

```
f <- as.vector(fast)
```

```
heatmapdata <- seqidentity(fast)

heatmapdata
```

```
##                XP_031992464.1 XP_017732256.1 XP_005566031.1 NP_006735.2
## XP_031992464.1          1.000          0.995          0.995       0.995
## XP_017732256.1          0.995          1.000          0.990       0.990
## XP_005566031.1          0.995          0.990          1.000       0.990
## NP_006735.2             0.995          0.990          0.990       1.000
## NP_001038960.1          0.995          0.990          0.990       1.000
## XP_003904062.1          1.000          0.995          0.995       0.995
##                NP_001038960.1 XP_003904062.1
## XP_031992464.1          0.995          1.000
## XP_017732256.1          0.990          0.995
## XP_005566031.1          0.990          0.995
## NP_006735.2             1.000          0.995
## NP_001038960.1          1.000          0.995
## XP_003904062.1          0.995          1.000
```

```
heatmap <- heatmap(heatmapdata, margins = c(6,6))
```



Q8. Top 3 unique hits for similar atomic resolution structures

```
#We can combine our sequences
conseq <- consensus(fast)
conseq
```

```
## $seq
##   [1] "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "M" "K" "W" "V" "W"
##  [19] "A" "L" "L" "L" "L" "A" "A" "L" "G" "S" "G" "R" "A" "E" "R" "D" "C" "R"
##  [37] "V" "S" "S" "F" "R" "V" "K" "E" "N" "F" "D" "K" "A" "R" "F" "S" "G" "T"
##  [55] "W" "Y" "A" "M" "A" "K" "K" "D" "P" "E" "G" "L" "F" "L" "Q" "D" "N" "I"
##  [73] "V" "A" "E" "F" "S" "V" "D" "E" "T" "G" "Q" "M" "S" "A" "T" "A" "K" "G"
##  [91] "R" "V" "R" "L" "L" "N" "N" "W" "D" "V" "C" "A" "D" "M" "V" "G" "T" "F"
## [109] "T" "D" "T" "E" "D" "P" "A" "K" "F" "K" "M" "K" "Y" "W" "G" "V" "A" "S"
## [127] "F" "L" "Q" "K" "G" "N" "D" "D" "H" "W" "I" "I" "D" "T" "D" "Y" "D" "T"
## [145] "Y" "A" "V" "Q" "Y" "S" "C" "R" "L" "L" "N" "L" "D" "G" "T" "C" "A" "D"
## [163] "S" "Y" "S" "F" "V" "F" "S" "R" "D" "P" "N" "G" "L" "P" "P" "E" "A" "Q"
## [181] "K" "I" "V" "R" "Q" "R" "Q" "E" "E" "L" "C" "L" "A" "R" "Q" "Y" "R" "L"
## [199] "I" "V" "H" "N" "G" "Y" "C" "D" "G" "R" "S" "E" "R" "N" "L" "L"
##
## $freq
##           1         2         3         4         5         6         7
## V 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000
## I 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000
## L 0.0000000 0.0000000 0.0000000 0.0000000 0.1666667 0.0000000 0.0000000
```

```
## M 0.1666667 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000
## F 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000
## W 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000
## Y 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000
## S 0.0000000 0.0000000 0.0000000 0.1666667 0.0000000 0.0000000 0.0000000
## T 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000
## N 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000
## Q 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.1666667
## H 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000
## K 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000
## R 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000
## D 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000
## E 0.0000000 0.1666667 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000
## A 0.0000000 0.0000000 0.1666667 0.0000000 0.0000000 0.0000000 0.0000000
## G 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000
## P 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.1666667 0.0000000
## C 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000
## - 0.8333333 0.8333333 0.8333333 0.8333333 0.8333333 0.8333333 0.8333333
## X 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000
##           8         9        10        11        12        13 14 15 16
## V 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000  0  0  0
## I 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000  0  0  0
## L 0.0000000 0.0000000 0.0000000 0.1666667 0.0000000 0.0000000  0  0  0
## M 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000  1  0  0
## F 0.0000000 0.0000000 0.1666667 0.0000000 0.0000000 0.0000000  0  0  0
## W 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000  0  0  1
## Y 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000  0  0  0
## S 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000  0  0  0
## T 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000  0  0  0
## N 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000  0  0  0
## Q 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000  0  0  0
## H 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000  0  0  0
## K 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.1666667  0  1  0
## R 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000  0  0  0
## D 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000  0  0  0
## E 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000  0  0  0
## A 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000  0  0  0
## G 0.1666667 0.1666667 0.0000000 0.0000000 0.1666667 0.0000000  0  0  0
## P 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000  0  0  0
## C 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000  0  0  0
## - 0.8333333 0.8333333 0.8333333 0.8333333 0.8333333 0.8333333  0  0  0
## X 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000  0  0  0
##          17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39
## V 0.8333333  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0
## I 0.0000000  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## L 0.0000000  0  0  1  1  1  1  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0
## M 0.1666667  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## F 0.0000000  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## W 0.0000000  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## Y 0.0000000  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## S 0.0000000  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  1  1
## T 0.0000000  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## N 0.0000000  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## Q 0.0000000  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
```

```
## H 0.0000000  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## K 0.0000000  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## R 0.0000000  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  1  0  0  1  0  0  0
## D 0.0000000  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0
## E 0.0000000  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0
## A 0.0000000  0  1  0  0  0  0  1  1  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0
## G 0.0000000  0  0  0  0  0  0  0  0  0  1  0  1  0  0  0  0  0  0  0  0  0  0  0
## P 0.0000000  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## C 0.0000000  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0
## - 0.0000000  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## X 0.0000000  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65
## V  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## I  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## L  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## M  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0
## F  1  0  0  0  0  0  1  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## W  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0
## Y  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0
## S  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0
## T  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0
## N  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## Q  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## H  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## K  0  0  0  1  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  1  1  0  0  0  0
## R  0  1  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## D  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0
## E  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0
## A  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  1  0  1  0  0  0  0  0  0
## G  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  1
## P  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0
## C  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## -  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## X  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91
## V  0  0  0  0  0  0  0  1  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0
## I  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## L  1  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## M  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0
## F  0  1  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## W  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## Y  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## S  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  1  0  0  0  0  0  0
## T  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  1  0  0  0  0
## N  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## Q  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0
## H  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## K  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0
## R  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1
## D  0  0  0  0  1  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0
## E  0  0  0  0  0  0  0  0  0  1  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0
## A  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  1  0  1  0  0  0
## G  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  1  0
## P  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
```

```
## C  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## -  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## X  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##    92 93 94 95 96 97 98 99 100 101 102 103 104 105 106 107 108 109 110 111 112
## V  1  0  0  0  0  0  0  0   1   0   0   0   0   1   0   0   0   0   0   0   0
## I  0  0  0  0  0  0  0  0   0   0   0   0   0   0   0   0   0   0   0   0   0
## L  0  0  1  1  0  0  0  0   0   0   0   0   0   0   0   0   0   0   0   0   0
## M  0  0  0  0  0  0  0  0   0   0   0   0   1   0   0   0   0   0   0   0   0
## F  0  0  0  0  0  0  0  0   0   0   0   0   0   0   0   0   1   0   0   0   0
## W  0  0  0  0  0  0  1  0   0   0   0   0   0   0   0   0   0   0   0   0   0
## Y  0  0  0  0  0  0  0  0   0   0   0   0   0   0   0   0   0   0   0   0   0
## S  0  0  0  0  0  0  0  0   0   0   0   0   0   0   0   0   0   0   0   0   0
## T  0  0  0  0  0  0  0  0   0   0   0   0   0   0   0   1   0   1   0   1   0
## N  0  0  0  0  1  1  0  0   0   0   0   0   0   0   0   0   0   0   0   0   0
## Q  0  0  0  0  0  0  0  0   0   0   0   0   0   0   0   0   0   0   0   0   0
## H  0  0  0  0  0  0  0  0   0   0   0   0   0   0   0   0   0   0   0   0   0
## K  0  0  0  0  0  0  0  0   0   0   0   0   0   0   0   0   0   0   0   0   0
## R  0  1  0  0  0  0  0  0   0   0   0   0   0   0   0   0   0   0   0   0   0
## D  0  0  0  0  0  0  0  1   0   0   0   1   0   0   0   0   0   0   1   0   0
## E  0  0  0  0  0  0  0  0   0   0   0   0   0   0   0   0   0   0   0   0   1
## A  0  0  0  0  0  0  0  0   0   0   1   0   0   0   0   0   0   0   0   0   0
## G  0  0  0  0  0  0  0  0   0   0   0   0   0   0   1   0   0   0   0   0   0
## P  0  0  0  0  0  0  0  0   0   0   0   0   0   0   0   0   0   0   0   0   0
## C  0  0  0  0  0  0  0  0   0   1   0   0   0   0   0   0   0   0   0   0   0
## -  0  0  0  0  0  0  0  0   0   0   0   0   0   0   0   0   0   0   0   0   0
## X  0  0  0  0  0  0  0  0   0   0   0   0   0   0   0   0   0   0   0   0   0
##    113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131
## V    0   0   0   0   0   0   0   0   0   0   0   1   0   0   0   0   0   0   0
## I    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
## L    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   1   0   0   0
## M    0   0   0   0   0   0   1   0   0   0   0   0   0   0   0   0   0   0   0
## F    0   0   0   0   1   0   0   0   0   0   0   0   0   0   1   0   0   0   0
## W    0   0   0   0   0   0   0   0   0   1   0   0   0   0   0   0   0   0   0
## Y    0   0   0   0   0   0   0   0   1   0   0   0   0   0   0   0   0   0   0
## S    0   0   0   0   0   0   0   0   0   0   0   0   0   1   0   0   0   0   0
## T    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
## N    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
## Q    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   1   0   0
## H    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
## K    0   0   0   1   0   1   0   1   0   0   0   0   0   0   0   0   0   1   0
## R    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
## D    1   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
## E    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
## A    0   0   1   0   0   0   0   0   0   0   0   0   1   0   0   0   0   0   0
## G    0   0   0   0   0   0   0   0   0   0   1   0   0   0   0   0   0   0   1
## P    0   1   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
## C    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
## -    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
## X    0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
##    132 133 134 135 136 137       138 139 140 141 142 143 144 145 146 147 148 149
## V    0   0   0   0   0   0 0.3333333   0   0   0   0   0   0   0   0   1   0   0
## I    0   0   0   0   0   1 0.6666667   0   0   0   0   0   0   0   0   0   0   0
## L    0   0   0   0   0   0 0.0000000   0   0   0   0   0   0   0   0   0   0   0
## M    0   0   0   0   0   0 0.0000000   0   0   0   0   0   0   0   0   0   0   0
```

```
## F   0   0   0   0   0   0 0.0000000   0   0   0   0   0   0   0   0   0   0   0
## W   0   0   0   0   1   0 0.0000000   0   0   0   0   0   0   0   0   0   0   0
## Y   0   0   0   0   0   0 0.0000000   0   0   0   1   0   0   1   0   0   0   1
## S   0   0   0   0   0   0 0.0000000   0   0   0   0   0   0   0   0   0   0   0
## T   0   0   0   0   0   0 0.0000000   0   1   0   0   0   1   0   0   0   0   0
## N   1   0   0   0   0   0 0.0000000   0   0   0   0   0   0   0   0   0   0   0
## Q   0   0   0   0   0   0 0.0000000   0   0   0   0   0   0   0   0   0   1   0
## H   0   0   0   1   0   0 0.0000000   0   0   0   0   0   0   0   0   0   0   0
## K   0   0   0   0   0   0 0.0000000   0   0   0   0   0   0   0   0   0   0   0
## R   0   0   0   0   0   0 0.0000000   0   0   0   0   0   0   0   0   0   0   0
## D   0   1   1   0   0   0 0.0000000   1   0   1   0   1   0   0   0   0   0   0
## E   0   0   0   0   0   0 0.0000000   0   0   0   0   0   0   0   0   0   0   0
## A   0   0   0   0   0   0 0.0000000   0   0   0   0   0   0   0   1   0   0   0
## G   0   0   0   0   0   0 0.0000000   0   0   0   0   0   0   0   0   0   0   0
## P   0   0   0   0   0   0 0.0000000   0   0   0   0   0   0   0   0   0   0   0
## C   0   0   0   0   0   0 0.0000000   0   0   0   0   0   0   0   0   0   0   0
## -   0   0   0   0   0   0 0.0000000   0   0   0   0   0   0   0   0   0   0   0
## X   0   0   0   0   0   0 0.0000000   0   0   0   0   0   0   0   0   0   0   0
##   150 151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168
## V   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   1   0
## I   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
## L   0   0   0   1   1   0   1   0   0   0   0   0   0   0   0   0   0   0   0
## M   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
## F   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   1   0   1
## W   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
## Y   0   0   0   0   0   0   0   0   0   0   0   0   0   0   1   0   0   0   0
## S   1   0   0   0   0   0   0   0   0   0   0   0   0   1   0   1   0   0   0
## T   0   0   0   0   0   0   0   0   0   1   0   0   0   0   0   0   0   0   0
## N   0   0   0   0   0   1   0   0   0   0   0   0   0   0   0   0   0   0   0
## Q   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
## H   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
## K   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
## R   0   0   1   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
## D   0   0   0   0   0   0   0   1   0   0   0   0   1   0   0   0   0   0   0
## E   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
## A   0   0   0   0   0   0   0   0   0   0   0   1   0   0   0   0   0   0   0
## G   0   0   0   0   0   0   0   0   1   0   0   0   0   0   0   0   0   0   0
## P   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
## C   0   1   0   0   0   0   0   0   0   0   1   0   0   0   0   0   0   0   0
## -   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
## X   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
##   169 170 171 172 173 174 175 176 177 178 179 180       181 182 183 184 185 186
## V   0   0   0   0   0   0   0   0   0   0   0   0 0.0000000   0   1   0   0   0
## I   0   0   0   0   0   0   0   0   0   0   0   0 0.0000000   1   0   0   0   0
## L   0   0   0   0   0   0   1   0   0   0   0   0 0.0000000   0   0   0   0   0
## M   0   0   0   0   0   0   0   0   0   0   0   0 0.0000000   0   0   0   0   0
## F   0   0   0   0   0   0   0   0   0   0   0   0 0.0000000   0   0   0   0   0
## W   0   0   0   0   0   0   0   0   0   0   0   0 0.0000000   0   0   0   0   0
## Y   0   0   0   0   0   0   0   0   0   0   0   0 0.0000000   0   0   0   0   0
## S   1   0   0   0   0   0   0   0   0   0   0   0 0.0000000   0   0   0   0   0
## T   0   0   0   0   0   0   0   0   0   0   0   0 0.0000000   0   0   0   0   0
## N   0   0   0   0   1   0   0   0   0   0   0   0 0.0000000   0   0   0   0   0
## Q   0   0   0   0   0   0   0   0   0   0   0   1 0.0000000   0   0   0   1   0
## H   0   0   0   0   0   0   0   0   0   0   0   0 0.0000000   0   0   0   0   0
```

```
## K   0   0   0   0   0   0   0   0   0   0   0   0 0.8333333   0   0   0   0   0
## R   0   1   0   0   0   0   0   0   0   0   0   0 0.1666667   0   0   1   0   1
## D   0   0   1   0   0   0   0   0   0   0   0   0 0.0000000   0   0   0   0   0
## E   0   0   0   0   0   0   0   0   0   1   0   0 0.0000000   0   0   0   0   0
## A   0   0   0   0   0   0   0   0   0   0   1   0 0.0000000   0   0   0   0   0
## G   0   0   0   0   0   1   0   0   0   0   0   0 0.0000000   0   0   0   0   0
## P   0   0   0   1   0   0   0   1   1   0   0   0 0.0000000   0   0   0   0   0
## C   0   0   0   0   0   0   0   0   0   0   0   0 0.0000000   0   0   0   0   0
## -   0   0   0   0   0   0   0   0   0   0   0   0 0.0000000   0   0   0   0   0
## X   0   0   0   0   0   0   0   0   0   0   0   0 0.0000000   0   0   0   0   0
##   187 188 189 190 191 192 193 194 195 196 197 198 199 200 201 202 203 204 205
## V   0   0   0   0   0   0   0   0   0   0   0   0   0   1   0   0   0   0   0
## I   0   0   0   0   0   0   0   0   0   0   0   0   1   0   0   0   0   0   0
## L   0   0   0   1   0   1   0   0   0   0   0   1   0   0   0   0   0   0   0
## M   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
## F   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
## W   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
## Y   0   0   0   0   0   0   0   0   0   1   0   0   0   0   0   0   0   1   0
## S   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
## T   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
## N   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   1   0   0   0
## Q   1   0   0   0   0   0   0   0   1   0   0   0   0   0   0   0   0   0   0
## H   0   0   0   0   0   0   0   0   0   0   0   0   0   0   1   0   0   0   0
## K   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
## R   0   0   0   0   0   0   0   1   0   0   1   0   0   0   0   0   0   0   0
## D   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
## E   0   1   1   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
## A   0   0   0   0   0   0   1   0   0   0   0   0   0   0   0   0   0   0   0
## G   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   1   0   0
## P   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
## C   0   0   0   0   1   0   0   0   0   0   0   0   0   0   0   0   0   0   1
## -   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
## X   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
##   206 207 208 209 210 211 212 213 214
## V   0   0   0   0   0   0   0   0   0
## I   0   0   0   0   0   0   0   0   0
## L   0   0   0   0   0   0   0   1   1
## M   0   0   0   0   0   0   0   0   0
## F   0   0   0   0   0   0   0   0   0
## W   0   0   0   0   0   0   0   0   0
## Y   0   0   0   0   0   0   0   0   0
## S   0   0   0   1   0   0   0   0   0
## T   0   0   0   0   0   0   0   0   0
## N   0   0   0   0   0   0   1   0   0
## Q   0   0   0   0   0   0   0   0   0
## H   0   0   0   0   0   0   0   0   0
## K   0   0   0   0   0   0   0   0   0
## R   0   0   1   0   0   1   0   0   0
## D   1   0   0   0   0   0   0   0   0
## E   0   0   0   0   1   0   0   0   0
## A   0   0   0   0   0   0   0   0   0
## G   0   1   0   0   0   0   0   0   0
## P   0   0   0   0   0   0   0   0   0
## C   0   0   0   0   0   0   0   0   0
```

```
## -   0   0   0   0   0   0   0   0   0
## X   0   0   0   0   0   0   0   0   0
##
## $seq.freq
##         1         2         3         4         5         6         7         8
## 0.1666667 0.1666667 0.1666667 0.1666667 0.1666667 0.1666667 0.1666667 0.1666667
##         9        10        11        12        13        14        15        16
## 0.1666667 0.1666667 0.1666667 0.1666667 0.1666667 1.0000000 1.0000000 1.0000000
##        17        18        19        20        21        22        23        24
## 0.8333333 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
##        25        26        27        28        29        30        31        32
## 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
##        33        34        35        36        37        38        39        40
## 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
##        41        42        43        44        45        46        47        48
## 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
##        49        50        51        52        53        54        55        56
## 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
##        57        58        59        60        61        62        63        64
## 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
##        65        66        67        68        69        70        71        72
## 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
##        73        74        75        76        77        78        79        80
## 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
##        81        82        83        84        85        86        87        88
## 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
##        89        90        91        92        93        94        95        96
## 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
##        97        98        99       100       101       102       103       104
## 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
##       105       106       107       108       109       110       111       112
## 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
##       113       114       115       116       117       118       119       120
## 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
##       121       122       123       124       125       126       127       128
## 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
##       129       130       131       132       133       134       135       136
## 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
##       137       138       139       140       141       142       143       144
## 1.0000000 0.6666667 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
##       145       146       147       148       149       150       151       152
## 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
##       153       154       155       156       157       158       159       160
## 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
##       161       162       163       164       165       166       167       168
## 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
##       169       170       171       172       173       174       175       176
## 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
##       177       178       179       180       181       182       183       184
## 1.0000000 1.0000000 1.0000000 1.0000000 0.8333333 1.0000000 1.0000000 1.0000000
##       185       186       187       188       189       190       191       192
## 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
##       193       194       195       196       197       198       199       200
## 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
```

```
##       201       202       203       204       205       206       207       208
## 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
##       209       210       211       212       213       214
## 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
##
## $cutoff
## [1] 0.6
```

```
conseq2 <- conseq$seq
conseq2
```

```
##   [1] "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "M" "K" "W" "V" "W"
##  [19] "A" "L" "L" "L" "L" "A" "A" "L" "G" "S" "G" "R" "A" "E" "R" "D" "C" "R"
##  [37] "V" "S" "S" "F" "R" "V" "K" "E" "N" "F" "D" "K" "A" "R" "F" "S" "G" "T"
##  [55] "W" "Y" "A" "M" "A" "K" "K" "D" "P" "E" "G" "L" "F" "L" "Q" "D" "N" "I"
##  [73] "V" "A" "E" "F" "S" "V" "D" "E" "T" "G" "Q" "M" "S" "A" "T" "A" "K" "G"
##  [91] "R" "V" "R" "L" "L" "N" "N" "W" "D" "V" "C" "A" "D" "M" "V" "G" "T" "F"
## [109] "T" "D" "T" "E" "D" "P" "A" "K" "F" "K" "M" "K" "Y" "W" "G" "V" "A" "S"
## [127] "F" "L" "Q" "K" "G" "N" "D" "D" "H" "W" "I" "I" "D" "T" "D" "Y" "D" "T"
## [145] "Y" "A" "V" "Q" "Y" "S" "C" "R" "L" "L" "N" "L" "D" "G" "T" "C" "A" "D"
## [163] "S" "Y" "S" "F" "V" "F" "S" "R" "D" "P" "N" "G" "L" "P" "P" "E" "A" "Q"
## [181] "K" "I" "V" "R" "Q" "R" "Q" "E" "E" "L" "C" "L" "A" "R" "Q" "Y" "R" "L"
## [199] "I" "V" "H" "N" "G" "Y" "C" "D" "G" "R" "S" "E" "R" "N" "L" "L"
```

```
blastResults <- blast.pdb(conseq2, database = "pdb")
```

```
##  Searching ... please wait (updates every 5 seconds) RID = UKWNNEA901R
##  ..
##  Reporting 103 hits
```

```
blastResults$hit.tbl
```

```
##          queryid subjectids identity alignmentlength mismatches gapopens
## 1   Query_327205     4O9S_A   98.387             186          3        0
## 2   Query_327205     3FMZ_A   98.387             186          3        0
## 3   Query_327205     6QBA_A   99.454             183          1        0
## 4   Query_327205     1JYD_A   99.451             182          1        0
## 5   Query_327205     1BRP_A   99.451             182          1        0
## 6   Query_327205     1JYJ_A   98.352             182          3        0
## 7   Query_327205     1QAB_E   97.778             180          4        0
## 8   Query_327205     2WQA_E   98.870             177          2        0
## 9   Query_327205     3BSZ_E   99.432             176          1        0
## 10  Query_327205     2WQ9_A   99.425             174          1        0
## 11  Query_327205     2WR6_A   98.857             175          2        0
## 12  Query_327205     1AQB_A   93.443             183         12        0
## 13  Query_327205     1HBQ_A   92.896             183         13        0
## 14  Query_327205     1ERB_A   92.896             183         13        0
## 15  Query_327205     1KT5_A   93.714             175         11        0
## 16  Query_327205     1RLB_E   93.103             174         12        0
## 17  Query_327205     1IIU_A   86.628             172         23        0
## 18  Query_327205     5EZ2_A   26.404             178        116        7
## 19  Query_327205     5F6Z_A   30.137             146         89        6
```

```
## 20  Query_327205     2HZQ_A   27.083          144         82         7
## 21  Query_327205     2NND_A   23.684          152         93         6
## 22  Query_327205     4R0B_A   26.168          107         68         4
## 23  Query_327205     1IW2_A   26.829          123         80         4
## 24  Query_327205     2QOS_C   25.439          114         77         3
## 25  Query_327205     2OVD_A   25.439          114         77         3
## 26  Query_327205     1GKA_B   28.767           73         49         2
## 27  Query_327205     2OVA_A   25.410          122         83         3
## 28  Query_327205     6GQZ_A   20.635          126         71         5
## 29  Query_327205     4ES7_A   22.314          121         66         5
## 30  Query_327205     4IAX_A   23.387          124         70         6
## 31  Query_327205     3KZA_A   29.907          107         58         6
## 32  Query_327205     3QKG_A   23.140          121         65         5
## 33  Query_327205     7L5M_A   32.653           49         33         0
## 34  Query_327205     1QWD_A   26.829           82         59         1
## 35  Query_327205     2ACO_A   26.829           82         59         1
## 36  Query_327205     3DSZ_A   22.222          126         69         6
## 37  Query_327205     6VRI_A   32.653           49         33         0
## 38  Query_327205     3MBT_A   26.829           82         59         1
## 39  Query_327205     7L5K_A   32.653           49         33         0
## 40  Query_327205     3OJY_C   24.561          114         78         3
## 41  Query_327205     1EW3_A   22.785          158         82         7
## 42  Query_327205     2RD7_C   24.561          114         78         3
## 43  Query_327205     4ORW_A   23.288          146         76         7
## 44  Query_327205     4ORR_A   23.288          146         76         7
## 45  Query_327205     1JZU_A   22.302          139         96         4
## 46  Query_327205     6UKK_A   26.829           82         59         1
## 47  Query_327205     5NGH_A   22.581          155         94         6
## 48  Query_327205     6UKL_A   32.653           49         33         0
## 49  Query_327205     4OS0_A   23.288          146         76         7
## 50  Query_327205     6UBO_A   26.829           82         59         1
## 51  Query_327205     1GKA_A   48.649           37         17         2
## 52  Query_327205     1GM6_A   21.512          172         94         9
## 53  Query_327205     2K23_A   21.368          117         77         3
## 54  Query_327205     4K6M_A   32.258           93         52         6
## 55  Query_327205     1S2P_A   48.649           37         17         2
## 56  Query_327205     4ALO_A   48.649           37         17         2
## 57  Query_327205     5MHH_A   19.841          126         72         5
## 58  Query_327205     1I4U_A   48.649           37         17         2
## 59  Query_327205     3DTQ_A   22.222          126         69         6
## 60  Query_327205     4MTP_A   31.818           66         39         3
## 61  Query_327205     4HDG_A   31.818           66         39         3
## 62  Query_327205     3NAP_C   30.337           89         57         2
## 63  Query_327205     1BSO_A   30.476          105         56         6
## 64  Query_327205     3GTN_A   32.653           49         33         0
## 65  Query_327205     1UZ2_X   30.476          105         56         6
## 66  Query_327205     1CJ5_A   30.476          105         56         6
## 67  Query_327205     4GH7_A   22.308          130         72         7
## 68  Query_327205     5X7Y_A   23.077          169         91         7
## 69  Query_327205     1YUP_A   29.907          107         58         6
## 70  Query_327205   7BH0_AAA   28.319          113         64         6
## 71  Query_327205     1Z24_A   28.440          109         70         4
## 72  Query_327205     4IAW_A   21.138          123         74         5
## 73  Query_327205     6QI7_A   29.524          105         57         6
```

```
## 74  Query_327205     3BX7_A   18.852             122         78        4
## 75  Query_327205     4NLI_A   29.524             105         57        6
## 76  Query_327205     4OMW_A   29.524             105         57        6
## 77  Query_327205     5NUM_A   29.204             113         63        6
## 78  Query_327205     6RWQ_A   28.319             113         64        6
## 79  Query_327205   7BGA_AAA   28.319             113         64        6
## 80  Query_327205     6NRE_A   22.727             154         86        6
## 81  Query_327205     5NUN_A   29.204             113         63        6
## 82  Query_327205     6RWR_A   28.319             113         64        6
## 83  Query_327205     5HTD_A   28.319             113         64        6
## 84  Query_327205     5NUJ_A   29.204             113         63        6
## 85  Query_327205     1B0O_A   29.524             105         57        6
## 86  Query_327205     5K06_A   29.524             105         57        6
## 87  Query_327205   7BF8_AAA   28.319             113         64        6
## 88  Query_327205     1BEB_A   29.524             105         57        6
## 89  Query_327205     3PH5_A   29.524             105         57        6
## 90  Query_327205     7LWC_A   29.524             105         57        6
## 91  Query_327205     6NKQ_A   29.524             105         57        6
## 92  Query_327205     6S8V_A   20.635             126         71        5
## 93  Query_327205     6QPD_A   29.524             105         57        6
## 94  Query_327205     5NUK_A   29.204             113         63        6
## 95  Query_327205     2L9C_A   18.125             160        118        4
## 96  Query_327205     5N47_A   20.635             126         71        6
## 97  Query_327205     2XST_A   22.523             111         72        3
## 98  Query_327205     6QPE_A   27.434             113         65        6
## 99  Query_327205     4OS8_A   22.603             146         77        7
## 100 Query_327205     4OS3_A   22.603             146         77        7
## 101 Query_327205     1QWK_A   40.000              55         29        1
## 102 Query_327205     1EPA_A   26.549             113         61        4
## 103 Query_327205     2GLE_A   27.907              43         25        1
##     q.start q.end s.start s.end    evalue bitscore positives  mlog.evalue
## 1        16   201      30   215 2.08e-139    389.0     98.92 319.32696003
## 2        16   201      27   212 2.37e-139    388.0     98.92 319.19643797
## 3        19   201       3   185 3.39e-138    384.0    100.00 316.53591291
## 4        19   200       2   183 1.19e-137    383.0    100.00 315.28020443
## 5        19   200       1   182 1.30e-137    383.0    100.00 315.19179348
## 6        19   200       2   183 1.15e-135    378.0     98.90 310.70922561
## 7        22   201       1   180 3.22e-133    372.0     98.33 305.07443601
## 8        18   194       1   177 4.35e-133    371.0     99.44 304.77364152
## 9        19   194       1   176 4.90e-133    371.0    100.00 304.65458216
## 10       19   192       1   174 1.58e-131    367.0    100.00 301.18122234
## 11       18   192       1   175 1.64e-131    367.0     99.43 301.14395094
## 12       19   201       1   183 4.48e-131    366.0     96.17 300.13902414
## 13       19   201       1   183 2.51e-130    364.0     97.27 298.41577934
## 14       19   201       1   183 3.41e-130    364.0     96.72 298.10934980
## 15       19   193       1   175 7.58e-125    350.0     96.57 285.79762342
## 16       19   192       1   174 1.03e-123    347.0     96.55 283.18840764
## 17       21   192       2   173 3.86e-116    328.0     94.77 265.74920360
## 18       23   198       7   171  4.94e-08     52.0     41.57  16.82331541
## 19       23   168       7   139  6.64e-08     51.2     43.84  16.52756878
## 20       29   166      11   137  1.82e-05     44.7     47.22  10.91408896
## 21       32   178      21   154  1.80e-04     42.0     42.76   8.62255371
## 22       28   132       4   101  1.00e-03     39.3     46.73   6.90775528
## 23       24   144      13   127  3.00e-03     38.1     45.53   5.80914299
```

```
## 24       24   136       4   110  1.10e-02     36.6     43.86   4.50986001
## 25       24   136      13   119  1.20e-02     36.6     43.86   4.42284863
## 26      124   196     105   174  2.50e-02     35.4     49.32   3.68887945
## 27       24   144      13   127  3.00e-02     35.4     42.62   3.50655790
## 28       24   136       9   118  5.80e-02     34.3     42.06   2.84731227
## 29       26   134      36   140  7.40e-02     34.3     42.98   2.60369019
## 30       24   136      13   122  8.40e-02     34.3     41.13   2.47693848
## 31       31   132       8   102  1.40e-01     33.1     42.99   1.96611286
## 32       26   134       9   113  1.60e-01     33.1     43.80   1.83258146
## 33       29    77      28    76  2.00e-01     32.0     53.06   1.60943791
## 34       29   110      30   110  2.20e-01     32.7     45.12   1.51412773
## 35       29   110      26   106  2.40e-01     32.7     45.12   1.42711636
## 36       24   136      13   122  2.60e-01     32.7     43.65   1.34707365
## 37       29    77      28    76  2.70e-01     31.6     53.06   1.30933332
## 38       29   110      12    92  3.20e-01     32.3     45.12   1.13943428
## 39       29    77      30    78  3.50e-01     32.3     53.06   1.04982212
## 40       24   136      13   119  3.70e-01     32.0     42.98   0.99425227
## 41       32   173       5   138  3.80e-01     32.0     39.87   0.96758403
## 42       24   136      15   121  4.00e-01     32.0     42.98   0.91629073
## 43        6   137      10   133  4.10e-01     32.0     43.84   0.89159812
## 44        6   137      10   133  4.30e-01     32.0     43.84   0.84397007
## 45       34   172       5   131  4.70e-01     31.6     38.85   0.75502258
## 46       29   110      30   110  4.90e-01     31.6     45.12   0.71334989
## 47       29   178       7   140  4.90e-01     31.6     41.94   0.71334989
## 48       29    77      28    76  5.20e-01     30.8     53.06   0.65392647
## 49        6   137      10   133  5.30e-01     31.6     43.84   0.63487827
## 50       29   110      30   110  5.60e-01     31.6     45.12   0.57981850
## 51      124   160     106   140  6.60e-01     31.2     62.16   0.41551544
## 52       29   189       9   150  6.70e-01     31.2     40.12   0.40047757
## 53       26   136      15   122  6.80e-01     31.2     42.74   0.38566248
## 54       13    98     260   348  7.00e-01     32.0     47.31   0.35667494
## 55      124   160     107   141  7.10e-01     31.2     62.16   0.34249031
## 56      124   160     107   141  7.10e-01     31.2     62.16   0.34249031
## 57       24   136      13   122  7.10e-01     31.2     42.86   0.34249031
## 58      124   160     107   141  7.50e-01     31.2     62.16   0.28768207
## 59       24   136      13   122  7.70e-01     31.2     42.86   0.26136476
## 60       35    98      16    77  9.00e-01     31.6     46.97   0.10536052
## 61       35    98      21    82  9.30e-01     31.6     46.97   0.07257069
## 62        8    91     106   194  9.30e-01     31.2     41.57   0.07257069
## 63       33   132      10   102  1.00e+00     30.8     40.95   0.00000000
## 64       72   120      72   120  1.10e+00     31.2     51.02  -0.09531018
## 65       33   132      10   102  1.10e+00     30.4     40.95  -0.09531018
## 66       33   132      10   102  1.20e+00     30.4     40.95  -0.18232156
## 67       24   140      13   126  1.30e+00     30.4     43.85  -0.26236426
## 68       14   171       1   141  1.30e+00     30.4     39.05  -0.26236426
## 69       31   132       8   102  1.40e+00     30.4     41.12  -0.33647224
## 70       25   132       2   102  1.80e+00     30.0     41.59  -0.58778666
## 71       32   138      17   119  1.90e+00     30.0     42.20  -0.64185389
## 72       24   136      13   122  2.00e+00     30.0     43.09  -0.69314718
## 73       33   132      10   102  2.00e+00     29.6     41.90  -0.69314718
## 74       24   136      13   122  2.00e+00     30.0     42.62  -0.69314718
## 75       33   132      10   102  2.10e+00     29.6     40.95  -0.74193734
## 76       33   132      10   102  2.10e+00     29.6     40.95  -0.74193734
## 77       25   132       2   102  2.30e+00     29.6     41.59  -0.83290912
```

```
## 78       25   132       2   102  2.40e+00     29.6     40.71  -0.87546874
## 79       25   132       2   102  2.40e+00     29.6     40.71  -0.87546874
## 80       29   171       8   139  2.40e+00     29.6     40.26  -0.87546874
## 81       25   132       2   102  2.60e+00     29.6     41.59  -0.95551145
## 82       25   132       2   102  2.80e+00     29.3     40.71  -1.02961942
## 83       25   132       2   102  3.10e+00     29.3     40.71  -1.13140211
## 84       25   132       2   102  3.30e+00     29.3     41.59  -1.19392247
## 85       33   132      10   102  3.40e+00     29.3     40.95  -1.22377543
## 86       33   132      11   103  3.40e+00     29.3     40.95  -1.22377543
## 87       25   132       2   102  3.40e+00     29.3     40.71  -1.22377543
## 88       33   132      10   102  3.40e+00     29.3     40.95  -1.22377543
## 89       33   132       9   101  3.40e+00     29.3     40.95  -1.22377543
## 90       33   132      10   102  3.70e+00     28.9     40.95  -1.30833282
## 91       33   132      26   118  3.70e+00     29.3     40.95  -1.30833282
## 92       24   136      13   122  3.80e+00     29.3     42.06  -1.33500107
## 93       33   132      10   102  3.90e+00     28.9     40.95  -1.36097655
## 94       25   132       2   102  4.00e+00     28.9     41.59  -1.38629436
## 95       23   182      14   160  5.30e+00     28.5     41.88  -1.66770682
## 96       24   136      13   122  5.30e+00     28.9     42.86  -1.66770682
## 97       29   136       8   107  6.20e+00     28.5     40.54  -1.82454929
## 98       25   132       2   102  6.90e+00     28.1     40.71  -1.93152141
## 99        6   137      10   133  7.00e+00     28.5     43.84  -1.94591015
## 100       6   137      10   133  7.00e+00     28.5     43.84  -1.94591015
## 101       7    57     248   302  7.60e+00     28.5     52.73  -2.02814825
## 102      29   134       2    99  8.10e+00     28.1     41.59  -2.09186406
## 103     122   158      14    56  9.60e+00     26.6     44.19  -2.26176310
##      pdb.id      acc
## 1   4O9S_A   4O9S_A
## 2   3FMZ_A   3FMZ_A
## 3   6QBA_A   6QBA_A
## 4   1JYD_A   1JYD_A
## 5   1BRP_A   1BRP_A
## 6   1JYJ_A   1JYJ_A
## 7   1QAB_E   1QAB_E
## 8   2WQA_E   2WQA_E
## 9   3BSZ_E   3BSZ_E
## 10  2WQ9_A   2WQ9_A
## 11  2WR6_A   2WR6_A
## 12  1AQB_A   1AQB_A
## 13  1HBQ_A   1HBQ_A
## 14  1ERB_A   1ERB_A
## 15  1KT5_A   1KT5_A
## 16  1RLB_E   1RLB_E
## 17  1IIU_A   1IIU_A
## 18  5EZ2_A   5EZ2_A
## 19  5F6Z_A   5F6Z_A
## 20  2HZQ_A   2HZQ_A
## 21  2NND_A   2NND_A
## 22  4R0B_A   4R0B_A
## 23  1IW2_A   1IW2_A
## 24  2QOS_C   2QOS_C
## 25  2OVD_A   2OVD_A
## 26  1GKA_B   1GKA_B
## 27  2OVA_A   2OVA_A
```

```
## 28  6GQZ_A   6GQZ_A
## 29  4ES7_A   4ES7_A
## 30  4IAX_A   4IAX_A
## 31  3KZA_A   3KZA_A
## 32  3QKG_A   3QKG_A
## 33  7L5M_A   7L5M_A
## 34  1QWD_A   1QWD_A
## 35  2ACO_A   2ACO_A
## 36  3DSZ_A   3DSZ_A
## 37  6VRI_A   6VRI_A
## 38  3MBT_A   3MBT_A
## 39  7L5K_A   7L5K_A
## 40  3OJY_C   3OJY_C
## 41  1EW3_A   1EW3_A
## 42  2RD7_C   2RD7_C
## 43  4ORW_A   4ORW_A
## 44  4ORR_A   4ORR_A
## 45  1JZU_A   1JZU_A
## 46  6UKK_A   6UKK_A
## 47  5NGH_A   5NGH_A
## 48  6UKL_A   6UKL_A
## 49  4OS0_A   4OS0_A
## 50  6UBO_A   6UBO_A
## 51  1GKA_A   1GKA_A
## 52  1GM6_A   1GM6_A
## 53  2K23_A   2K23_A
## 54  4K6M_A   4K6M_A
## 55  1S2P_A   1S2P_A
## 56  4ALO_A   4ALO_A
## 57  5MHH_A   5MHH_A
## 58  1I4U_A   1I4U_A
## 59  3DTQ_A   3DTQ_A
## 60  4MTP_A   4MTP_A
## 61  4HDG_A   4HDG_A
## 62  3NAP_C   3NAP_C
## 63  1BSO_A   1BSO_A
## 64  3GTN_A   3GTN_A
## 65  1UZ2_X   1UZ2_X
## 66  1CJ5_A   1CJ5_A
## 67  4GH7_A   4GH7_A
## 68  5X7Y_A   5X7Y_A
## 69  1YUP_A   1YUP_A
## 70  7BH0_a 7BH0_AAA
## 71  1Z24_A   1Z24_A
## 72  4IAW_A   4IAW_A
## 73  6QI7_A   6QI7_A
## 74  3BX7_A   3BX7_A
## 75  4NLI_A   4NLI_A
## 76  4OMW_A   4OMW_A
## 77  5NUM_A   5NUM_A
## 78  6RWQ_A   6RWQ_A
## 79  7BGA_a 7BGA_AAA
## 80  6NRE_A   6NRE_A
## 81  5NUN_A   5NUN_A
```

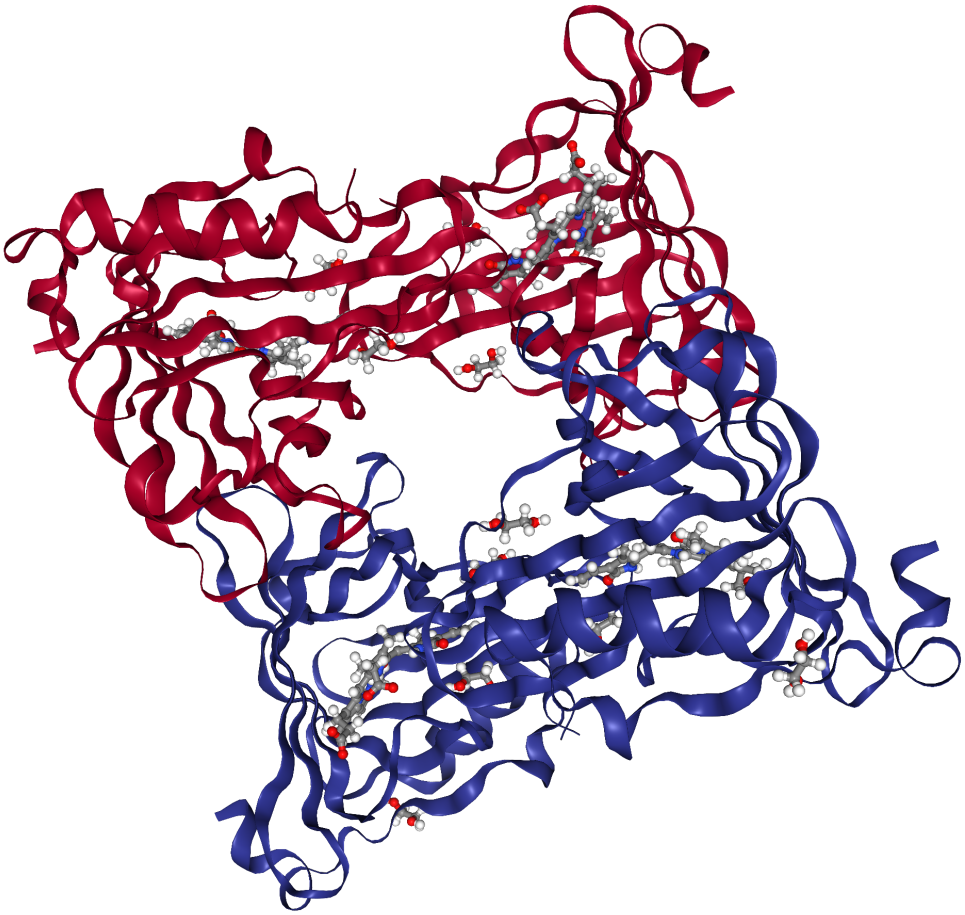
```
## 82  6RWR_A   6RWR_A
## 83  5HTD_A   5HTD_A
## 84  5NUJ_A   5NUJ_A
## 85  1BOO_A   1BOO_A
## 86  5K06_A   5K06_A
## 87  7BF8_a 7BF8_AAA
## 88  1BEB_A   1BEB_A
## 89  3PH5_A   3PH5_A
## 90  7LWC_A   7LWC_A
## 91  6NKQ_A   6NKQ_A
## 92  6S8V_A   6S8V_A
## 93  6QPD_A   6QPD_A
## 94  5NUK_A   5NUK_A
## 95  2L9C_A   2L9C_A
## 96  5N47_A   5N47_A
## 97  2XST_A   2XST_A
## 98  6QPE_A   6QPE_A
## 99  4OS8_A   4OS8_A
## 100 4OS3_A   4OS3_A
## 101 1QWK_A   1QWK_A
## 102 1EPA_A   1EPA_A
## 103 2GLE_A   2GLE_A
```

1. Chain A, Sandercyanin Fluorescent Protein (5EZ2_A) Evalue : 4.94e-08 ; sequence identity : 26.404 Source organism : Sander vitreus (walleye) experimentalTechnique : X-Ray Diffraction Resolution : 1.85A
2. Chain A, Apolipoprotein D (2HZQ_A) Evalue : 1.82e-05 ; sequence identity : 27.083 Source organism : Homo sapiens (human) experimentalTechnique : X-Ray Diffraction Resolution : 1.8A
3. Chain A, Major urinary protein 2 (PDB : 2NND_A) Evalue : 1.80e-04 ; sequence identity : 23.684 Source organism : Mus musculus (house mouse) experimentalTechnique : X-Ray Diffraction Resolution : 1.6A

Q9. Generate molecular figure

I will use NGL viewer online

This is for our first result, 5EZ2_A



This structure only had a 26.404% sequence identity compared to our "novel" protein, it is likely that this structure is not very similar to our novel one. There may be a few conserved residues and base strucutre parts but as a whole will be different from ours.

> Q10. Perform a "Target" search of chEMBEL w/ our novel sequence. Are there any Target Associated Assays and ligand efficiency data reported that may be useful starting points for exploring potential inhibition of your novel protein?

This was our initail results page for our chEMBEL search https://www.ebi.ac.uk/chembl/g/#search_results/assays/query=MKWVWALLLLAALGSGRAERDCRVSSFRVKENFDKARFSGTWYAMAKKDPEGLFLQDNIVAE20MSATAKGRVRLLNNWDVCADMVGTFTDTEDPAKFKMKYWGVASFLQKGNDDHWIIDTDYDTYAVQYSCRL%20LNLDGTCADSYSFVFSRDPNGLPPEAQRIVRQRQEELCLARQYRLIVHNGYCDGRSERNLL

There were 1,383,553 assays found, and when searched for target associated, there were 20,346 results from those assays. CHEMBL3881277 (https://www.ebi.ac.uk/chembl/assay_report_card/CHEMBL3881277/)

looked interesting as it had target levels of decrease in heme oxygenase protein expression labels that could be related to our retinol binding protein.

CHEMBL1293256 has to do with the ligand thrombopoietin, and had assays measuring its potency and functionality. https://www.ebi.ac.uk/chembl/target_report_card/CHEMBL1293256/